Findings of the WMT Shared Task on Machine Translation Using Terminologies

Md Mahfuz ibn Alam[1], **Ivana Kvapilíková**[5], **Antonios Anastasopoulos**[1], **Laurent Besacier**[2], **Georgiana Dinu**[3], **Marcello Federico**[3], **Matthias Gallé**[2], **Philipp Koehn**[4], **Vassilina Nikoulina**[2], **Kweon Woo Jung**[2]
[1]Department of Computer Science, George Mason University [2]NAVER
[3]AWS [4]Facebook / Johns Hopkins University [5]Charles University
{malam21, antonis}@gmu.edu

Abstract

Language domains that require very careful use of terminology are abundant and reflect a significant part of the translation industry. In this work we introduce a benchmark for evaluating the quality and consistency of terminology translation, focusing on the medical (and COVID-19 specifically) domain for five language pairs: English to French, Chinese, Russian, and Korean, as well as Czech to German. We report the descriptions and results of the participating systems, commenting on the need for further research efforts towards both more adequate handling of terminologies as well as towards a proper formulation and evaluation of the task.

1 Introduction

Language domains that require very careful use of terminology are abundant. The need to adequately translate within such domains is undeniable, as shown by e.g. the different WMT shared tasks on biomedical translation.

More interestingly, as the abundance of research on domain adaptation shows, such language domains are (a) not adequately covered by existing data and models, while (b) new (or "surge") domains arise and models need to be adapted, often with significant downstream implications: consider the new COVID-19 domain and the large efforts for translation of critical information regarding pandemic handling and infection prevention strategies.

In the case of newly developed domains, while parallel data are hard to come by, it is fairly straightforward to create word- or phrase-level terminologies, which can be used to guide professional translators and ensure both accuracy and consistency.

This shared task[1] replicated such a scenario, and invited participants to explore methods to incorporate terminologies into either the training or the inference process, in order to improve both the accuracy and consistency of MT systems on a new domain.

2 Shared Task Details

The shared task focused on five language pairs, with systems evaluated on:

- English to French
- English to Chinese
- English to Russian
- English to Korean
- Czech to German

The last three language pairs were "surprise" language pairs. This shared task construction follows a three-phase approach to ensure the generalizability of the findings, inspired by other multilingual shared tasks (Vylomova et al., 2020). In this setting, only part of the evaluation language pairs (or languages) are revealed from the beginning (the **Development Phase**). In this elongate period (a couple of months), the participants are provided with data in some language pairs to *develop* their methods. The second phase is the **Generalization phase**, which is a short time period (two to three weeks in this task's case), in which additional (surprise) language settings are revealed, only giving the shared task participants enough time to deploy a system, as opposed to allowing them enough time to also perform extensive optimization on the datasets. The final stage is the **Evaluation phase**, in which the test data are released and the methods are evaluated on these held-out data.

The goal of this 3-stage approach (with both development and surprise language pairs) is to avoid approaches that overfit on language selection, and instead evaluate the more realistic scenario of needing to tackle the new domain in a new language in a limited amount of time. The surprise language pairs were announced 3 weeks before the start of the evaluation campaign.

The organizers provided training/development

[1] http://statmt.org/wmt21/terminology-task.html

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 652–663
November 10–11, 2021.

data and terminologies for the above language pairs. Test sets were released at the beginning of the evaluation period. The participating teams were invited to participate in any or all of the language pairs.

2.1 Data

Training The shared task primarily focused on a constrained submission setting, in which the participants could only use any parallel or monolingual data listed in previous versions of WMT shared tasks to train their systems. Some pre-trained systems listed at the shared task announcement (mBERT, XLM, XLM-R, mBART, mT5, M2M100) were also allowed, but should be disclosed by the participants. We note that the training data allowed come from a "general" domain, as opposed to e.g. highly specialized biomedical data, which in theory should be more helpful for this setting.

Terminologies The shared task focused on adapting MT systems to the health domain in general, with a particular interest in the surge COVID-19 domain.

The terminologies for the English to French, Chinese, Russian, and Korean language pairs were taken from the publicly available TICO-19 project (Anastasopoulos et al., 2020), a multi-organizational project that created data to aid translators and evaluate MT systems on the COVID-19 domain. The terminologies were created by linguists at Google and Facebook in consultation with domain experts, providing translations for about 600 terms in each language. The terminologies are publicly available.[2]

The Czech-German medical terminology was generated automatically from Wikipedia. We considered all Wikipedia titles corresponding to the category *Health care* or to one of its subcategories, and all titles linked from the text. The list of (sub)categories was manually filtered to only include relevant articles. We treated all page titles as terms and relied on the Wikipedia language links to provide their translations. Furthermore, we used redirection links to obtain synonyms of both source and target terms.

For all terminologies, we truecased the terms using a pretrained truecaser and manually checked the results. The Czech-German terminology was eventually further reduced to only include terms which occurred in the EMEA medical corpus.

[2] https://tico-19.github.io/

Development and Test The development and test data for French, Chinese, and Russian were taken from the publicly available TICO-19 evaluation data. The organizers additionally created Korean translations of the English source-side sentences, which will be made available as part of the original TICO-19 datasets.[3]

The primary source of the Czech-German development and test data is the EMEA[4] parallel corpus of the European Medicines Agency. We cleaned it using the Moses tools, searched for terms and their translations and tagged the occurrences. The surface forms used for the search were collected from a corpus of in-domain Wikipedia articles which includes links to the lemmatized Wikipedia titles/terms next to their inflected forms. Target options were retrieved from the terminology and enriched with surface forms. Out of all sentences with terms, we selected around 3.5k sentences for the dev set and 1.1k for the test set. The development and test sets were tagged automatically but the test set was manually corrected to get rid of the artifacts caused by the automatic generation.

2.2 Ensuring Terminology Consistency on the Evaluation Datasets

It is worth noting that, originally, none of the development and test data were created under the constraints imposed by the specific terminologies we use. As such, we needed to ensure that the data 'complied' with the terminologies in order to guarantee a meaningful, accurate, and fair to the participants evaluation of the shared task's research questions.

The TICO-19 project created the evaluation dataset independently of the terminologies.[5] In our preliminary analysis, we first searched for all terminology terms on the English side of the parallel data, also searching over the lemmatized versions of English sentences. The choice of starting from the English side is due to two reasons: (a) it reflects the actual translation direction the data was created with and that we evaluate on, (b) it reduces the rate of possible false negative/positive term matches due to the lack of morphological complexity of English.

[3] The data are freely available here: https://tico-19.github.io/.

[4] https://opus.nlpl.eu/EMEA.php

[5] Although we note that the dataset went through an independent quality assurance process and several correction iterations, if required.

Example 1 (ID: Wikipedia_handpicked_4:1709)	
Source:	after blowing your nose , <term, src='coughing', tgt='tousser'> coughing </term> or <term, src='sneezing', tgt='éternuer'> sneezing </term> .
Translation:	après s ' être mouché ou avoir toussé / éternué ;
Annotation 1:	<term, src='coughing', tgt='tousser'> Label: c) variation_correct
Annotation 2:	<term, src='sneezing', tgt='éternuer'> Label: c) variation_correct
Tagged translation:	après s ' être mouché ou avoir <term, src='coughing'> toussé </term> / <term, src='sneezing'> éternué </term>;
Term-compl. transl.:	N/A
Example 2 (ID: Wikipedia_handpicked_4:1703)	
Source:	people can also become <term, src='infected', tgt='infecté'> infected </term> with <term, src='respiratory disease', tgt='maladie respiratoire'> respiratory diseases </term> such as <term, src='influenza', tgt='grippe'> influenza </term> or the <term, src='common cold', tgt='rhume'> common cold </term> , for example , if they do not wash their hands before <term, src='touch', tgt='toucher'> touching </term> their eyes , nose , or mouth (i . e . , mucous membranes) .
Translation:	il est possible de contracter des maladies respiratoires telles que la grippe ou le rhume , par exemple , en omettant de se laver les mains avant de se toucher les yeux , le nez ou la bouche (c . - à - d . les muqueuses) .
Annotation 1:	<term, src='infected', tgt='infecté'> Label: e) not_used
Annotation 2:	<term, src='respiratory disease', tgt='maladie respiratoire'> Label: c) variation_correct
Annotation 3:	<term, src='influenza', tgt='grippe'> Label: b) exact_match_correct
Annotation 4:	<term, src='common cold', tgt='rhume'> Label: b) exact_match_correct
Annotation 5:	<term, src='touch', tgt='toucher'> Label: b) exact_match_correct
Tagged translation:	il est possible de contracter des <term, src='respiratory disease'> maladies respiratoires</term> telles que la grippe ou le rhume , par exemple , en omettant de se laver les mains avant de se toucher les yeux , le nez ou la bouche (c . - à - d . les muqueuses) .
Term-compl. transl.:	il est possible d'etre <term, src= infected> infecté </term> avec des <term, src='respiratory disease'> maladies respiratoires</term> telles que grippe ou le rhume , par exemple , en omettant de se laver les mains avant de se toucher les yeux , le nez ou la bouche (c . - à - d . les muqueuses) .
Example 3 (ID: CMU_1:77	
Source:	I have hay <term, src='fever', tgt='fièvre'> fever </term> though too
Translation:	mais j ' ai le rhume des foins aussi
Annotation 1:	<term, src='fever', tgt='fièvre'> Label: a) does_not_apply
Tagged translation:	N/A
Term-compl. transl.:	N/A
Example 4 (ID: Wikipedia_handpicked_1:1311	
Source:	the strongest <term, src='self quarantine', tgt='auto - quarantaine'> self - <term, src='quarantine', tgt='quarantaine'> quarantine </term> </term> instructions have been issued to those in high risk groups .
Translation:	les instructions de quarantaine individuelle les plus strictes ont été données aux personnes des groupes les plus à risque .
Annotation 1:	<term, src='self quarantine', tgt='auto - quarantaine'> Label: e) not_used
Annotation 2:	<term, src='quarantine', tgt='quarantaine'> Label: b) exact_match_correct
Tagged translation:	N/A
Term-compl. transl.:	les instructions d' <term, src='self quarantine'> auto - quarantaine </term> les plus strictes ont été données aux personnes des groupes les plus à risque

Table 1: Examples (from English-French TICO-19) of expected annotations that ensure that the evaluation datasets are compliant with the terminologies. ('Term-compl. transl.' == 'terminology-compliant translation').

Having the source-side terms identified, we assume all of them should be translated according to the terminology. We then search the target side (both original and lemmatized) for the translation required by the terminology, and created a tag on the source-side term if we found an exact match. Last, we showed all sentences to professional translators, who were instructed to produce three types of annotations, for each source-side term. The first is a **label** describing whether (a) the automatically-annotated source-side term should not be translated by the terminology i.e. it is not really a term, (b) the tagged exact match is correct, (c) the translation is compliant with the terminology even though there is not an exact match, (d) the tagged exact match is incorrect, or (e) the source term translation is applicable in the context, but not used. The second annotation is a **tagged translation** for any terms labeled as (a), (c), or (d), denoting exactly which part of the target-side corresponds to the source-side term. The third annotation is a **tagged terminology-compliant translation**, where if any

source-side terms are labeled with (d) or (e), we ask the translators to rephrase the target side in order to make it compliant with the terminology.

Table 1 shows example sentences from the dataset, along with their expected annotations from the translators. Below we provide the exact instructions given to the annotators, which also reference the same examples.

[begin annotation instructions]
About: This task is about determining if a translation is compliant with a terminology data base and perform inline annotations on the translations to mark the terms used.
Annotators receive: Source side input, together with approximate terminology matches on the source side.
Annotators return: For each term match, please annotate a Label:

(a) `does_not_apply`: The terminology is not applicable in the context because of wrong meaning on the source side (Example 3). Please use *a) if you think the translation should not comply with the terminology matched*, irrespective of whether the translation uses it or not.
(b) `exact_match_correct`: The term translation is found exactly as is in the target and its usage is correct (it fits the context and agrees grammatically with the sentence). (Example 2)
(c) `variation_correct`: The translation is compliant with the terminology, however the term translation appears in a different form in the target (Examples 1 and 2). If only part of the term was preserved, use this label if this partial term is sufficient and completely preserves the meaning. Please use *b) or c) if you think the translation is compliant with the terminology*.
(d) `incorrect`: The term is found in the target, as an exact match or as a variant, but it is used incorrectly, either semantically or grammatically: e.g. the term use does not convey the required meaning, there is a wrong inflection or other grammatical disagreement.
(e) `not_used`: The term translation is applicable in the context, but not used (Example 2, 4). Make this only for clear omissions: everything else should be variation (correct or incorrect variation) Please use *d) or e) if you think the translation is not compliant with the terminology, but it should*.

Tagged translation: For any terms that are labeled as a), c) or d) please add inline markup to identify the fragments of the translation that they match.
For each source sentence, please generate a **Tagged Terminology-compliant translation**: if any of d) to e) apply to any term in the sentence, meaning the translation is not compliant with the terminology for at least one term, please provide an alternate translation that is compliant with the terminology w.r.t all the terms in the sentence. If there is no acceptable translation that would use the expected target term then you should annotate the target with a) does_not_apply. If all terms in sentence match a) b) or c), leave this empty.
[/end annotation instructions]

Through this process, we ended up modifying 284 (9.25%), 251 (8.17%), 450 (14.65%), and 809 (26.34%) sentences in the French, Chinese, Russian, and Korean datasets respectively, in order to make them terminology-compliant. Last, the Czech-German terminologies were directly derived from the parallel data hence they implicitly directly reflect the underlying data, so there was no need for the aforementioned process.

2.3 Evaluation

The evaluation of the shared task used several metrics, focusing on both translation accuracy and terminological consistency.

- Translation accuracy was evaluated with standard reference-based MT metrics (BLEU, chrF, BERTscore, COMET). In light of recent work (Kocmi et al., 2021), we rank systems according to the COMET metric.

- we also performed terminology-targeted evaluation (to evaluate for consistency). We use the metrics outlined by Alam et al. (2021), namely exact-match term accuracy, 1-TERm score, and window overlap accuracy. We rank systems according to term exact-match accuracy.

Briefly, the lemmatized exact-match term accuracy is an accuracy score that searches for exact term translation matches (of the terminology required output) over either the lemmatized or the original hypothesis. The window overlap accuracy identifies the translation of the term, and then

scores its context, to measure how well a translated term is placed in the hypothesis. Last, the 1-TERm score is a modification of the TER metric (Snover et al., 2006), biased to assign higher edit cost weights for words belonging to a term (and then simply reversed so that a higher score is better). We refer the reader to Alam et al. (2021) for further discussion of the metrics and supporting arguments for their use.

Last, we evaluate whether differences between systems are statistically significant using paired bootstrap resampling (Koehn, 2004), over sentence-level COMET and exact-match accuracy scores. Based on this information we cluster statistically-insignificantly-different (i.e. similarly performing) systems when we produce their final rankings.

Winning submissions will be the ones that are Pareto-optimal along the two evaluation metrics that a good but also terminology-compliant system should maximize: exact-match accuracy (which captures terminology consistency) and COMET (which captures general translation quality). As such, there is the possibility that each language pair will have multiple winning submissions.

3 Participants and System Descriptions

We received a total of 43 submissions from 9 teams. Below we provide a short description of each submission.

CUNI (Jon et al., 2021b) Authors competed on `En-Fr` language pair. The terminology constrains are inserted as done in (Jon et al., 2021a). The target translation of specific terms is appended to the source sentence as a suffix and separated by a special token (if multiple constraints occur for a single sentence, an additional token separator is added). In order to have more training data of this form, synthetic constraints are added by sampling random token subsequences from the target sentence and appending them to the source sentence as described earlier. Note that since no modification is done on target side of the parallel data, no post-processing of the MT output is needed. As NMT systems trained from this pre-processed data sometimes fail to generate inflection in the translation output, terminology tokens appended to the source are lemmatized for both training and inference which brings improvements over the different shared task metrics.

Huawei (HW-TSC) (Wang et al., 2021b) Authors submitted output of an unconstrained system to `En-Zh` language pair. They train a Transformer *big* architecture on both out-of-domain and in-domain (biomedical) data. Parallel data in biomedical domain is augmented using more resources from TAUS[6] and back-translation of monolingual in-domain data is also applied. For the terminology shared-task, authors applied the system created for the biomedical translation shared task (described in (Wang et al., 2021b)) without any specific adaptation except appending the terminology dictionary to the end of training data. No separate paper was submitted for the terminology task.

Kakao Enterprises (KEP) (Bak et al., 2021) Authors submitted to `En-Fr`, `En-Zh`, `En-Kr`, `Cz-De`. A detailed data cleaning is performed, removing between 6% and 14% of the data. In-domain data is back-translated (only for `En-Fr` and `En-Kr`) and is selected by a a combination of keywords spotting and domain similarity, measured as perplexity of an in-domain language model. A first model is obtained by adding to that synthetic language pairs obtained by verbalizing the terminology database. The only language pair were this verbalization does not yield improvement is `Cz-De`, whose terminology was automatically constructed. Models obtained in this manner were submitted for `En-Zh`, `En-Kr`, `Cz-De`.

For `En-Fr` additional techniques are used: as those obtained the highest COMET score we detail them there. The final system for that language pair is trained inspired by techniques from (Bergmanis and Pinnis, 2021a; Dinu et al., 2019), but without modifying the model architecture. The source data is modified by adding immediately after a source term the corresponding target *lemma*, separated by special tokens. The model is pre-trained on randomly selected verbs and nouns, and fine-tuned using the terminology ontology. Interestingly, the pre-trained model - while improving Exact Match with respect to the baselines - degrades all other metrics. That degradation is however recovered and even improved when fine-tuning. For `En-Ko` and `Cs-De` ensemble models were used.

Lingua Custodia (LC) (Ailem et al., 2021a) The team participated in `En-Fr`, `En-Ru` and `En-Zh` tasks. They build on top of (Ailem et al., 2021b) by inserting the terminology as constraints in the

[6] https://md.taus.net/corona

source sentence. Such constraints represent special tags around the detected source term followed by the target term from the terminology, the original source term is masked. Presence of such constraints at training encourages the model to copy the correct term translation. In case where multiple translations are proposed by the terminology, the one which is present in the target sentence is chosen at training time. At inference time the translation is selected at random. In order to enforce learning signal, the team enriched parallel data with back-translation of monolingual data that contains terminology. Authors show that the proposed method allows to improve significantly for standard MT evaluation metrics, as well as terminology oriented metrics (Alam et al., 2021) over the standard baseline without terminological constraints.

PROMT (Molchanov et al., 2021) The team submitted two systems (`En-Fr` and `En-Ru`), both of which are transformer models implemented on MarianMT (Junczys-Dowmunt et al., 2018). The first approach uses a rule-based system (SmartMT) to modify the neural system's output, which extracts rules only for noun phrases. If the desired output of a source term is not found in the NMT output, the rule-based system identifies the term's current translation and its morphological analysis (case and number) in order to substitute it with the terminology-provided translation in the desired inflection. The second approach is an adaptation of (Dinu et al., 2019) to MarianNMT toolkit. Each source terminological term is followed by its translation using special tokens to signal these terminological entries in the text (and impose a soft-constraint to the translation system). Model is retrained from such pre-processed data. Data augmentation is also performed to create more synthetic data with terminology markup. Both approaches are rather close in performance.

SPECTRANS (Ballier et al., 2021) This team sumbitted to `En-Fr` language pair. They experimented with 2 open source NMT toolkits JoyeNMT (Kreutzer et al., 2019) and OpenNMT (Klein et al., 2017). After the first experiments with Europarl they retained OpenNMT which gave better performance. Their best runs were trained on CommonCrawl augmented with terminological data. They provided qualitative analysis of terminology-related translations and discuss the limitations of the terminologies provided for the task.

SYSTRAN (Pham et al., 2021) This participant submitted to `En-Fr` language pair and proposed two methods to incorporate terminology. The first approach, based on (Michon et al., 2020), replaces source and target terminological terms by placeholders including a unique identifier plus morphological information (masculine/feminine and singular/plural). In a variant of this method, the source terminology word form is also incorporated in the source stream. At training time, NMT model is learnt on such pre-processed data and a post-processing step recovers the word tokens from the placeholders after inference. The second approach (which lead to better performance) consists in learning a copy behaviour for terminological tokens at training time: terminology translations are inserted in the source sentence either by appending the target term (its surface or lemma form) to its source version, or by directly replacing the original term with the target one. A NMT system is trained on such pre-processed data and no post-process for recovering terminology tokens is needed at inference as target side of parallel data remains untouched. For both approaches however, a grammatical error correction is applied to the MT hypotheses in order to limit morphology errors. The impact of such post-processing on BLEU is positive, although small.

TermMind (Wang et al., 2021a) The team submitted to `En-Zh` task. Similar to (Ailem et al., 2021a) they build on top of (Ailem et al., 2021b) by inserting terminological constraints in the training data. In the case where multiple translations are available they augment source sentence with all possible translations (which is different from (Ailem et al., 2021a) who kept only one translation). In order to strengthen the learning signal participants extend given terminologies with biphrases extracted from parallel data and integrate the constraints for those biphrases as well. Finally, they used backtranslation, fine tuning on pseudo in-domain data and ensembling to strengthen the baseline model. Ensembling methods seem to lead to the best results.

TILDE (Bergmanis and Pinnis, 2021c) The team participated to `En-Fr`, `En-Ru` and `Cz-De` language pairs. They focused primarily on terminology filtering, outlining several notable shortcomings of the Shared Task's terminologies, most of which are due to the use of terminologies intended

for human translators (as opposed to terminologies created specifically for integration with MT systems). They devise two strategies for selecting among multiple target candidates for a source term, finding that an alignment-based technique outperforms the option of always selecting the first terminology entry. The MT systems are transformer-based using MarianMT, also integrating the method of Bergmanis and Pinnis (2021b) for incorporating terminology constraints in a soft manner.

4 Results and Discussion

The results and rankings for English-French are listed in Table 2 and for English-Chinese in Table 3. The results for the surprise language pairs are in Table 4 for English-Russian and Table 5 for English-Korean and Czech-German.

In the English-French translation task, there are two winning submissions. Two ProMT submissions ranked first according to exact-match accuracy (along with a CUNI submission), but the ProMT.soft submission is statistically significantly better than the other two with respect to COMET, hence it is one of the winning submission. The second winning submission is the one by KEP, which ranks first according to COMET, but also according to 1-TER, which indicates that it might strike a good balance between general translation quality and term consistency.

In the English-Chinese translation task there is a single winning submission, the one by TermMind (system 2), which ranks first according to both metrics. We note that another submission (HW-TSC) is statistically significantly better than all submissions in all metrics except for 1-TERm, but this submission is an unconstrained one, and hence it is excluded from the rankings.

In English-Russian the ProMT submission ProMT.soft is the clear winner, ranking as the single best system according to exact-match accuracy, as well as one of the two best systems according to COMET. Interestingly, the other system that ranks first according to COMET (ProMT.smartnd.v2) ranks first according to 1-TERm score, but also *last* according to exact-match accuracy, denoting perhaps an orthogonality between the goals of terminological consistency and general translation quality, where prioritizing one over the other leads to performance drops along the other dimension.

Last, the submissions by KEP are the winning ones for English-Korean and Czech-German. For the former language pair it was the only submitted system (see discussion on potential reasons), while for Czech-German it ranked for best system according to exact match accuracy with the other submission (by TildeMT), but was significantly better according to COMET. Although TildeMT used a more sophisticated approach to the terminology translation, the KEP team had a stronger baseline and used ensembling which significantly increased both general translation quality and the term accuracy.

4.1 General Quality

It was pointed out by Bergmanis and Pinnis (2021c) that a majority of terms from the terminologies were represented in the training corpora, which could lead to an underestimation of the importance of terminology in the metrics. The results show that using terminology constrains leads to an improvement over the baselines trained without it, but the effect would be more substantial if the training corpora were filtered to exclude sentences with terms.

Perhaps a future iteration of the shared task could include an explicitly novel domain, although how well such a domain indeed exists or is even possible in the age of big data where our models can be trained on a large part of the Internet is debatable. An alternative is to carefully filter the training corpora to remove sentences with the terms, to create a truly challenging domain adaptation with terminologies setting.

4.2 Terminology Consistency

The discussion of the Shared Task taught us that narrow terminology with unambiguous translations is more suitable for terminology-focused machine translation than a broader and more universal terminology with several target options. Unlike human translators who naturally choose from translation alternatives, it is difficult for a MT system to filter out noisy or inappropriate word forms. While a narrow terminology can ensure a proper and exact translation of terms, e.g. when translating a lecture with several special terms known in advance, we believe that a broad terminology can serve for more general domain adaptation using existing lexical resources. We note that several participating teams highlighted this issue, e.g. Ballier et al. (2021); Bergmanis and Pinnis (2021c).

The TICO terminologies in a few cases included additional comments aimed at translators who are

English-French	Rankings		Terminology-focused				Translation Quality	
	accoding to		Exact-Match	Window Overlap		1-TERm		BLEU
System	ex-m. acc.	COMET	Accuracy	(2)	(3)	Score	COMET	(truecased)
ProMT.soft	1-3	3	**0.974**	**0.359**	**0.352**	0.625	0.752	47.69
ProMT.smartnd	1-3	4-5	0.966	0.357	0.348	0.626	0.746	47.89
CUNI-Primary_not_scored	1-3	6-10	0.967	0.342	0.334	0.601	0.732	46.92
KEP	4-6	1	0.950	0.34	0.337	**0.632**	**0.781**	**49.60**
CUNI-Primary_lemm	4-6	6-10	0.946	0.34	0.332		0.729	46.80
CUNI-Contr._not_scored	4-6	12-18	0.950	0.33	0.331	0.588	0.693	45.48
SYSTRAN-app+_corr	7-17	2	0.934	0.355	0.349	0.631	0.766	48.87
SYSTRAN-app_corr	7-17	6-10	0.938	0.283	0.297	0.614	0.729	45.81
SYSTRAN-mrk_corr	7-17	6-10	0.938	0.283	0.297	0.614	0.729	45.81
SYSTRAN-mrk+_corr	7-17	6-10	0.938	0.283	0.297	0.614	0.729	45.81
TildeMT	7-17	11	0.939	0.329	0.322	0.593	0.706	45.04
CUNI-Contr._sf_choices	7-17	12-18	0.923	0.313	0.310	0.557	0.682	42.72
LinguaCustodia-Sys1	7-17	12-18	0.920	0.343	0.336	0.595	0.677	44.49
LinguaCustodia-Sys2_new	7-17	12-18	0.919	0.344	0.335	0.598	0.681	44.90
LinguaCustodia-Sys2	7-17	12-18	0.919	0.345	0.334	0.591	0.676	44.21
CUNI-Contrastive_sf	7-17	12-18	0.918	0.321	0.317		0.684	44.08
CUNI-Contr._lemm_choices	7-17	12-18	0.913	0.323	0.317	0.567	0.678	43.78
ProMT.baseline	18	4-5	0.898	0.33	0.331	0.624	0.745	47.50
SPECTRANS3-CC-fr_en	19	19	0.871	0.296	0.296	0.507	0.596	40.02
SPECTRANS	20	20	0.795	0.27	0.267	0.495	0.296	34.93
SPECTRANS_2	21	21	0.640	0.248	0.241	0.480	0.212	33.59

Table 2: English-French results. The systems are ranked and clustered according to exact-match accuracy (secondarily according to COMET) based on statistical significance tests. We **highlight** the best score per metric.

English-Chinese	Rankings		Terminology-focused				Translation Quality	
	accoding to		Exact-Match	Window Overlap Acc.		1-TERm		BLEU
System	ex-m. acc.	COMET	Accuracy	(2)	(3)	Score	COMET	(truecased)
*HW-TSC**	*1**	*1**	***0.886***	***0.282***	***0.285***	*0.514*	***0.716***	***40.73***
TermMind-sys2	2	2	0.856	0.27	0.274	**0.534**	0.709	40.47
LinguaCustodia - Sys1-v2	3-6	4	0.828	0.225	0.227	0.438	0.643	29.61
LinguaCustodia - Sys1	3-6	5-7	0.829	0.223	0.225	0.437	0.637	29.16
LinguaCustodia - Sys2	3-6	5-7	0.829	0.222	0.225	0.433	0.635	28.92
LinguaCustodia - Sys1-v3	3-6	5-7	0.828	0.241	0.244	0.472	0.641	33.73
TermMind	7-8	3	0.668	0.22	0.227	0.513	0.696	37.51
KEP	7-8	8	0.645	0.18	0.187	0.249	0.229	27.12

Table 3: English-Chinese results. The systems are ranked and clustered according to exact-match accuracy based on statistical significance tests. We **highlight** the best score per metric. *: unrestricted system.

English-Russian	Rankings		Terminology-focused				Translation Quality	
	accoding to		Exact-Match	Window Overlap Acc.		1-TERm		BLEU
System	ex-m. acc.	COMET	Accuracy	(2)	(3)	Score	COMET	(truecased)
ProMT.soft	1	1-2	**0.909**	**0.254**	**0.255**	0.482	0.631	31.06
ProMT.smartnd.v1	2-5	3	0.857	0.25	0.250	0.482	0.624	31.52
LinguaCustodia - Sys1	2-5	5-8	0.854	0.24	0.249	0.472	0.598	28.84
LinguaCustodia - Sys1-v2	2-5	5-8	0.849	0.24	0.247	0.473	0.600	28.81
TildeMT-v2	2-5	9-10	0.863	0.22	0.226	0.457	0.550	28.14
LinguaCustodia - Sys2-v2	6-7	5-8	0.849	0.247	0.248	0.474	0.604	29.13
LinguaCustodia - Sys2	6-7	5-8	0.847	0.242	0.244	0.471	0.601	28.97
ProMT.baseline	8-9	4	0.823	0.24	0.241	0.481	0.620	31.49
TildeMT	8-9	9-10	0.817	0.21	0.219	0.456	0.548	28.16
ProMT.smartnd.v2	10	1-2	0.788	0.243	0.241	**0.487**	**0.634**	**31.92**

Table 4: English-Russian results. The systems are ranked and clustered according to exact-match accuracy (and secondarily according to COMET) based on statistical significance tests. We **highlight** the best score per metric.

directly looking at them, as opposed to the format that terminologies aimed at machines would use. We will take this into account in future iterations of the shared task – it is worth noting, though, that if most available terminologies are designed for human translators, it should probably be up to the

Language Pair	System	Rankings accoding to ex-m. acc.	COMET	Terminology-focused Exact-Match Accuracy	Window Overlap Acc. (2)	(3)	1-TERm Score	Translation Quality COMET	BLEU (truecased)
English-Korean	KEP	1	1	**0.569**	**0.067**	**0.065**	**0.251**	**0.581**	**16.52**
Czech-German	KEP	1-2	1	0.866	**0.428**	**0.424**	**0.474**	**0.694**	**34.10**
	TildeMT	1-2	2	**0.871**	0.390	0.385	0.434	0.641	30.01

Table 5: English-Korean and Czech-German results. The systems are ranked and clustered according to exact-match accuracy based on statistical significance tests. We **highlight** the best score achieved per metric.

NLP/ML/MT practitioners to figure out how to best use the existing data, rather than demanding new, dedicated resources. Similarly, when compiling the Czech-German terminology, we aimed at creating a universal lexicon of medical terms with a wide coverage. Many terms have multiple translations and we used the Wikipedia redirection links as a proxy for synonyms. Unfortunately, they became a source of noise because not all redirects are synonyms and not all synonyms are appropriate in every context. We tackled the former by semi-automatic filtering and left the latter up to the candidate translation engine to select the version of the word appropriate for the given context. Unfortunately, some problematic terms remained even in the final version of the terminology, as pointed out by Bergmanis and Pinnis (2021c).

4.3 Development vs Surprise Language Pairs

The participants had significantly more time to develop systems for English-French and English-Chinese, as opposed to the other three surprise language pairs. This is reflected partly on the total submitted systems in each language pair, where English-Korean and Czech-German received only 1 and 2 submissions respectively. We hypothesize that another explanation for this lies in the much more low-resource setting of these two language pairs, which generally tend to lead to lower quality systems, which might in turn discourage the participants.

A second potential explanation could lie in the general cohort of participants, which is largely comprised of teams from industry (the only exception is the CUNI team that is an academic one). Perhaps the two low-resource language pairs are simply translation directions that the participating institutions are less interested in – which we take as an indication for the importance of including such less-researched, low-resource, under-served language pairs in future iterations of this shared task, to encourage research in languages and language pairs beyond those with the most obvious commercial value.

4.4 Czech-German Analysis

We believe that even with the automatically generated resources this task provided an important insight into translation of terms between two linguistically different and morphologically rich languages such as German and Czech.

When analyzing the results, we focused on the phenomenon of nominal compounding in German. A natural translation of terms into German often results in a compound of a term and a general word, e.g. *Hormonproduktion* (production of hormones), or two terms, e.g. *Plasmaprotein* (plasma protein). Compounding is an important aspect of terminology-based translation to German that the model should have the capacity to create compounds from terminology entries.

The automatic metrics favor translations into two separate words, even though a compound is often more natural. We analyzed how candidate translations handled concrete cases; see Table 6 for an example. Out of 262 sentences with this phenomenon in the reference, the correct compound word was generated in 112 and 133 cases by the TildeMT and KEP systems, respectively. Both systems generate compounds from terms, although the former was trained with terminology constraints and the latter only saw the terms during explicit training on the terminology entries.

5 Related Work

Phrase-based statistical MT systems (Koehn et al., 2003) allowed for fine-grained control over the system's output by design, e.g. by incorporating domain-specific dictionaries into the phrase table, or by forcing translation choices for certain words or phrases. On the other hand, the currently state-of-the-art approach of neural machine translation (NMT) does not inherently allow for such control over the system's output. Some approaches

SRC	Mozkové metastázy vykazovaly nekonsistentní nebo žádnou fluorescenci.	...krvácení do svalů nebo hematom.
TGT	Hirnmetastasen zeigten inkonsistente oder keine Fluoreszenz.	...Muskelblutung oder Hämatom.
TildeMT	Zerebrale Metastasen zeigten eine inkonsistente oder keine Fluoreszenz.	...Blutungen in den Muskeln oder Hämatom
KEP	Hirnmetastasen zeigten eine inkonsistente oder keine Fluoreszenz.	...Muskelblutung oder Hämatom.

Table 6: Examples of term compounding in German where candidates handle term translation differently.

incorporate dictionaries through interpolation of the decoder's probability with a lexical probability based on source-side attention matches (Arthur et al., 2016). Perhaps the most common paradigm is *constrained decoding* (Hokamp and Liu, 2017; Anderson et al., 2017; Post and Vilar, 2018, *inter alia*), where the terminology matches are presented as hard constraints that the beam search must satisfy.

Constrained decoding is not without disadvantages: it can be computationally expensive and it is often brittle when applied in realistic conditions (Dinu et al., 2019). To this end, some works (Dinu et al., 2019; Bergmanis and Pinnis, 2021b; Exel et al., 2020; Niehues, 2021) introduced approaches where the terminological constraints are provided as input to the NMT as additional annotations inline with the source sentence. As such, these can be considered as "soft" constraints, as there is no guarantee that the NMT system will indeed produce an output containing them.

In any case, the best practice for incorporating terminological constraints in NMT is both underresearched and still not settled yet, especially in the case of morphologically rich languages, underlying the need for this shared task.

6 Conclusion

We presented the results of the first edition of the WMT21 shared task on MT using Terminologies. For the purposes of the task we created new evaluation datasets, annotated by professional translators for their terminology consistency, based on the TICO-19 data for English to French, Chinese, Russian, and Korean, as well as a dataset for Czech-German based on the EMEA corpus.

The Shared Task received 43 submissions from 9 teams, 8 from industry and 1 from academia, underscoring the general applicability of our focus problem ('how best can we use a terminology in MT?') on real-world settings. Most submissions add soft or hard constraints on the source side that the MT learns to handle, as proposed in (Dinu et al., 2019), but other novel approaches include terminology filtering for selecting between multiple options provided by the terminology, or replacing terms with placeholders to be inserted after the MT has produced the output. We devised multiple terminology-targeted metrics and evaluated systems along both these metrics as well as general translation quality. In most cases we find that, encouragingly, one does not necessarily have to sacrifice general translation quality for terminology compliance, as long as the terminology is of adequate standards.

In future iterations of the Shared Task, we will take into account the distinction between terminologies created for humans (which are abundant) and terminologies created specifically for MT systems which need to be created, and have different requirements/specifications that the former. In addition, we will attempt to consider a new domain, rather than focusing again on the biomedical domain and specifically COVID-19 (although this is a great example of a "surge" domain that immediately required that translation providers and MT engines adapt in order to handle translations of large volumes of text in this novel domain).

Acknowledgements

The organizers want to first thank the participants for their submissions and their constructive feedback during and after the Shared Task, which made the evaluation more robust. In addition, we are thankful to NAVER for creating the English-Korean translations and to Appen for quality assurance on them. The Czech-German track was organized with the support of the grant 1050119 of the Charles University Grant Agency. Last, we are thankful to Amazon, Tanya Badeka and Margo Lynch for the creation of the terminology-compliant translations of the evaluation datasets. Anastasopoulos is generously supported by NSF grant IIS-2125466.

References

Melissa Ailem, Jinghsu Liu, and Raheel Qader. 2021a. Lingua custodia's participation at the wmt 2021 machine translation using terminologies shared task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021b. Encouraging neural machine translation to satisfy terminology constraints. *CoRR*, abs/2106.03730.

Md Mahfuz Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021. On the evaluation of machine translation for terminology consistency. arXiv:2106.11891.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. Tico-19: the translation initiative for covid-19. In *NLP COVID-19 Workshop*, Online.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Yunju Bak, Jimin Sun, Jay Kim, Sungwon Lyu, and Changmin Lee. 2021. Kakao enterprise's wmt21 machine translation using terminologiestask submission. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Nicolas Ballier, Dahn Cho, Bilal Faye, Zong-You Ke, hanna Martikainen, Mojca Pecman, Jean-Baptiste Yunés, Guillaume Wisniewski, Lichao Zhu, and Maria Zimina-Poirot. 2021. The spectrans system description for the wmt21 terminology task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021a. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021b. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021c. Dynamic terminology integration for covid-19 and other emerging domains. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Josef Jon, João Paulo Aires, Dusan Varis, and Ondřej Bojar. 2021a. End-to-end lexically constrained machine translation for morphologically rich languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4019–4033, Online. Association for Computational Linguistics.

Josef Jon, Michal Novák, João Paulo Aires, Dušan Variš, and Ondrej Bojar. 2021b. Cuni systems for wmt21: Terminology translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. arXiv:1804.00344.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. arXiv:2107.10821.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Philipp Koehn, Franz J Och, and Daniel Marcu. 2003. Statistical phrase-based translation. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. *CoRR*, abs/1907.12484.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexander Molchanov, Vladislav Kovalenko, and Fedor Bykov. 2021. Promt systems for wmt21 terminology translation task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Jan Niehues. 2021. Continuous learning in neural machine translation using bilingual dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online. Association for Computational Linguistics.

MinhQuang Pham, Antoine Senellart, Dan Berrebbi, Josep Crego, and Jean Senellart. 2021. Systran @ wmt 2021: Terminology task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, volume 200. Citeseer.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Ke Wang, Shuqin Gu, Boxing Chen, Yu Zhao, Weihua Luo, and Yuqi Zhang. 2021a. Termmind: Alibaba's submission to the wmt21 machine translation using terminologies shared task. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Weixuan Wang, Wei Peng, Xupeng Meng, and Qun Liu. 2021b. Huawei aarc's submissions to the wmt21 biomedical translation task: Domain adaption from a practical perspective. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.